Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities

Year: 2023 | Citations: 103 | Authors: Maximilian Mozes, Xuanli He, Bennett Kleinberg, L. D. Griffin

---

Abstract

Spurred by the recent rapid increase in the development and distribution of large language models (LLMs) across industry and academia, much recent work has drawn attention to safety- and security-related threats and vulnerabilities of LLMs, including in the context of potentially criminal activities. Specifically, it has been shown that LLMs can be misused for fraud, impersonation, and the generation of malware; while other authors have considered the more general problem of AI alignment. It is important that developers and practitioners alike are aware of security-related problems with such models. In this paper, we provide an overview of existing - predominantly scientific - efforts on identifying and mitigating threats and vulnerabilities arising from LLMs. We present a taxonomy describing the relationship between threats caused by the generative capabilities of LLMs, prevention measures intended to address such threats, and vulnerabilities arising from imperfect prevention measures. With our work, we hope to raise awareness of the limitations of LLMs in light of such security concerns, among both experienced developers and novel users of such technologies.